High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding

Hannah Marienwald, Jean-Baptiste Fermanian, Gilles Blanchard

High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding

Hannah Marienwald* Jean-Baptiste Fermanian† Gilles Blanchard‡

Abstract

We propose an improved estimator for the multi-task averaging problem, whose goal is the joint estimation of the means of multiple distributions using separate, independent data sets. The naive approach is to take the empirical mean of each data set individually, whereas the proposed method exploits similarities between tasks, without any related information being known in advance. First, for each data set, similar or neighboring means are determined from the data by multiple testing. Then each naive estimator is shrunk towards the local average of its neighbors. We prove theoretically that this approach provides a reduction in mean squared error. This improvement can be significant when the dimension of the input space is large, demonstrating a "blessing of dimensionality" phenomenon. An application of this approach is the estimation of multiple kernel mean embeddings, which plays an important role in many modern applications. The theoretical results are verified on artificial and real world data.

1 INTRODUCTION

The estimation of means from i.i.d. data is arguably one of the oldest and most classical problems in statistics. In this work we consider the problem of estimating *multiple* means $\mu_1, \dots, \mu_B$ of probability distributions $\mathbb{P}_1, \dots, \mathbb{P}_B$, over a common space $\mathcal{X} = \mathbb{R}^d$ (or possibly a real Hilbert space $\mathcal{H}$). We assume that for each individual distribution $\mathbb{P}_i$, we observe an i.i.d. data set $X_\bullet^{(i)}$ of size N_i, and that these data sets have been collected independently from each other.

In the rest of the paper, we will call each such data set $X_\bullet^{(i)}$ a *bag*. Mathematically, our model is thus

$$\begin{cases} X_\bullet^{(i)} := (X_k^{(i)})_{1 \le k \le N_i} \overset{i.i.d.}{\sim} \mathbb{P}_i, \ 1 \le i \le B; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent}, \end{cases} \tag{1}$$

where $\mathbb{P}_1, \dots, \mathbb{P}_B$ are square integrable distributions on $\mathbb{R}^d$ which we call *tasks*, and our goal is the estimation of their means

$$\mu_i := \mathbb{E}_{X \sim \mathbb{P}_b}[X] \in \mathbb{R}^d, \ 1 \le i \le B. \tag{2}$$

Given an estimate $\widehat{\mu}_i$ of μ_i, we will be interested in its squared error $\|\widehat{\mu}_i - \mu_i\|^2$, and aim at controlling it either with high probability or in average (mean squared error, MSE):

$$\mathrm{MSE}(i, \widehat{\mu}_i) := \mathbb{E}\big[\|\widehat{\mu}_i - \mu_i\|^2\big];$$

this error can be considered either individually for each task $\mathbb{P}_i$ or averaged over all tasks.

*Universität Potsdam, Potsdam, Germany, and Technische Universität Berlin, Berlin, Germany.
†École Normale Supérieure de Rennes, Rennes, France
‡Université Paris-Saclay, CNRS, Inria, Laboratoire de Mathématiques d'Orsay, 91405, Orsay, France.

This problem is also known as multi-task averaging (MTA) (Feldman et al., 2014), an instance of the multi-task learning (MTL) problem. Prior work on MTL showed that learning multiple tasks jointly yields better performance compared to individual single task solutions (Caruana, 1997; Evgeniou et al., 2005; Feldman et al., 2014).

In this paper we adapt the idea of joint estimation to the multi-task averaging problem and will show that we can take advantage of some unknown *structure* in the set of tasks to improve the estimation. Here, by individual estimation we mean that our natural baseline is the naive estimator (NE) given by the simple empirical mean:

$$\widehat{\mu}_i^{\mathrm{NE}} := \frac{1}{N_i}\sum_{k=1}^{N_i} X_k^{(i)}; \quad \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) = \frac{1}{N_i}\operatorname{Tr}\Sigma_i, \tag{3}$$

where Σ_i is the covariance matrix of $\mathbb{P}_i$.

Our motivation for considering this setting is the growing number of large databases taking the above form, where independent bags corresponding to different but conceptually similar distributions are available; for example, one can think of i as an index for a large number of individuals, for each of which a number of observations (assumed to be sampled from an individual-specific distribution) are available, say medical records, or online activity information collected by some governmental or corporate mass spying device.

While estimating means in such databases is of interest of its own, a particularly important motivation to consider this setting is that of Kernel Mean Embedding (KME), a technique enjoying sustained attention in the statistical and machine learning community since its introduction in the seminal paper of Smola et al. (2007); see Muandet et al. (2017) for an overview. The KME methodology is used in a large number of applications, e.g. two sample testing (Gretton et al., 2012), goodness-of-fit (Chwialkowski et al., 2016), multiple instance or distributional learning for both supervised (Muandet et al., 2012; Szabó et al., 2016) as well as unsupervised learning (Jegelka et al., 2009), to name just a few.

The core principle of KME is to represent the distribution $\mathbb{P}_Z$ of a random variable Z via the mean of $X = \phi(Z)$, where ϕ is a rich enough feature mapping from the input space $\mathcal{Z}$ to a (reproducing kernel) Hilbert space $\mathcal{H}$. In practice, it is assumed that we have an i.i.d. bag $(Z_k)_{1\le k\le N}$ from $\mathbb{P}$, which is used to estimate its KME. Here we are interested again in the situation where a large number of independent data sets from different distributions are available, and we want to estimate their KMEs jointly. This is, therefore, an instance of the model (1), once we set $\mathcal{X} := \mathcal{H}$ and $X_k^{(i)} := \phi(Z_k^{(i)})$.

1.1 Relation to Previous Work

The fact that the naive estimator (3) can be improved upon when multiple, real-valued means are to be estimated simultaneously, has a long history in mathematical statistics. More precisely, let us introduce the following isotropic Gaussian setting:

$$\mathbb{P}_i = \mathcal{N}(\mu_i, I_i);\ N_i = N, \qquad 1 \le i \le B, \tag{GI}$$

on which we will come back in the sequel.

As shown in Stein (1956), for $B = 1$ with $d \ge 3$ the naive estimator is inadmissible, i.e. there exists a strictly better estimator, with a lower MSE for any true mean vector μ_1. An explicit example of a better estimator is given by the celebrated *James-Stein* estimator (JSE) (James and Stein, 1961), which shrinks adaptively the naive estimator towards $\mathbf{0}$, or more generally, towards an a priori fixed vector ν_0.

The MTA problem was introduced by Feldman et al. (2014), who proposed an approach which regularizes the estimation such that similar tasks shall have similar means as well. However, they assumed the pairwise task similarity to be given, which is unfeasible in most practical applications. In addition to our own approach, we will also introduce a variation of theirs, suitable for the KME framework, that *estimates* the task similarity instead of assuming it to be known. Martínez-Rego and Pontil (2013) proposed a method based on spectral clustering of the tasks and applying Feldman et al. (2014)'s method separately on each cluster, but without theoretical analysis.

Variations of the JSE can be shown to yield possible improvements over the NE in more general situations as well (see Fathi et al., 2020 for recent results in non-Gaussian settings). This has also been exploited for KME in Muandet et al. (2016), where a Stein-type estimator in kernel space was shown to generally improve over naive KME estimation. To the best of our knowledge, no shrinkage estimator for KME explicitly designed for or taking advantage of the MTA setting exists.

In the remainder of this work we will proceed as follows. Section 2 introduces the basic idea of the approach and starts with a general discussion. We will expose in Section 3 a theoretical analysis proving that the presented method improves upon the naive estimation in terms of squared error, possibly by a large factor. The general theoretical results will be discussed explicitly for the Gaussian setting (Sec. 3.3) and in the KME framework (Sec. 3.4). The approach is then tested for the KME setting on artificial and real world data in Section 4. All proofs are found in the appendix Sections A to F, Appendix G gives a detailed description of the estimators compared in the experiments, and Appendix H presents additional numerical results in the Gaussian setting.

2 METHOD

The basic idea of our approach is to improve the estimation of a mean of a task by basing its estimation not on its own bag alone, but concatenating the samples from all bags it is *sufficiently similar* to. Since in most practical applications task similarity is not known, we will propose a statistical test that assesses task relatedness based on the given data.

2.1 Overview of the Approach

In the remainder of the paper we will use the notation $[\![n]\!] := \{1, \dots, n\}$. For convenience of exposition, assume the (GI) setting. In this case, the naive estimators all have the same MSE, $\overline{\sigma}^2 := d/N$. Fix a particular task (reindexed $i = 0$) with mean μ_0 that we wish to estimate, and assume for now we are given the *side information* that for some constant $\tau > 0$, it holds $\Delta_{0i}^2 := \|\mu_0 - \mu_i\|^2 \leq \tau\overline{\sigma}^2$ for some "neighbor tasks" $i \in [\![V]\!]$ (a subset of the larger set of B tasks within range $\tau\overline{\sigma}^2$ to μ_0, reindexed for convenience). Consider the estimator $\widetilde{\mu}_0$ obtained by a simple average of neighbor naive estimators, $\widetilde{\mu}_0 = \frac{1}{V+1}\sum_{i=0}^{V}\widehat{\mu}_i^{\mathrm{NE}}$. We can bound via usual bias-variance decomposition, independence of the bags and convexity of the squared norm:

$$\mathrm{MSE}(0, \widetilde{\mu}_0) \;=\; \left\| \frac{1}{V+1}\sum_{i=1}^{V}(\mu_0 - \mu_i) \right\|^2 + \frac{\overline{\sigma}^2}{V+1} \;\leq\; \overline{\sigma}^2\frac{(1+V\tau)}{V+1}. \tag{4}$$

Thus, the above bound guarantees that $\widetilde{\mu}_0$ improves over $\widehat{\mu}_0^{\mathrm{NE}}$ whenever $\tau < 1$, and leads to a relative improvement of order $\max(\tau, V^{-1})$.

In practice, we *don't* have *any* a priori side information on the configuration of the means. A simple idea is, therefore, to estimate the quantities Δ_{0i}^2 from the data by an estimator $\widehat{\Delta}_{0i}^2$ and select only those bags for which $\widehat{\Delta}_{0i}^2 \leq \widetilde{\tau}\overline{\sigma}^2$. This is in a nutshell the principle of our proposed method.

The deceptive simplicity of the above idea might be met with some deserved skepticism. One might expect that the typical estimation error of $\widehat{\Delta}^2_{0i}$ would be of the same order as the MSE of the naive estimators. Consequently, we could at best guarantee with high probability a bound of $\Delta^2_{0i} \lesssim \overline{\sigma}^2$ for the estimated neighbor tasks, i.e. $\tau \approx 1$, which does not lead to any substantial theoretical improvement when using (4). The reason why the above criticism is pessimistic, even in the worst case, is the role of the dimension d. From high-dimensional statistics, it is known that the rate of *testing* for $\Delta^2_{0i} = 0$, i.e. the minimum ρ^2 such that a statistical test can detect $\Delta^2_{0i} \geq \rho^2$ with probability close to 1, is faster than the rate of *estimation*, $\rho^2 \simeq \sqrt{d}/N = \overline{\sigma}^2/\sqrt{d}$ (see e.g. Baraud, 2002; Blanchard et al., 2018). Thus, we can reliably determine neighbor tasks with $\tau \approx 1/\sqrt{d}$. Based on (4), we can hope again for an improvement of order up to $\mathcal{O}(1/\sqrt{d})$ over NE, which is significant even for a moderately large dimension. In the rest of the paper, we develop the idea sketched here more precisely and illustrate its consequences on KME by numerical experiments. The message we want to convey is that the *curse* of higher dimensional data with its effect on MSE can be to a limit mitigated by a *relative blessing* because we can take advantage of neighboring tasks more efficiently.

2.2 Proposed Approach

Denote $\overline{\sigma}^2_i = \mathrm{MSE}(i, \widehat{\mu}^{\mathrm{NE}}_i), i \in [\![B]\!]$. Introduce the following notation: $\Delta_{ij} := \|\mu_i - \mu_j\|$. In general, our approach assumes that we have at hand a family of tests $(T_{ij})_{1 \leq i,j \leq B}$ for the null hypotheses $H^0_{ij} : \Delta^2_{ij} > \tau \overline{\sigma}^2_i$ against the alternatives $H^1_{ij} : \Delta^2_{ij} \leq \tau' \overline{\sigma}^2_i$, for $0 \leq \tau' < \tau$. The exact form of the tests will be discussed later for specific settings.

We denote the set of detected neighbors of task $i \in [\![B]\!]$ as $V_i := \{j : T_{ij} = 1, j \in [\![B]\!]\}$; we can safely assume $T_{ii} = 1$ so that that $i \in V_i$ always holds and $|V_i| \geq 1$. We will also denote $V_i^* = V_i \setminus \{i\}$. For $\gamma \in [0, 1]$, define the modified estimator

$$\widetilde{\mu}_i := \gamma \widehat{\mu}^{\mathrm{NE}}_i + \frac{(1-\gamma)}{|V_i|} \sum_{j \in V_i} \widehat{\mu}^{\mathrm{NE}}_j, \tag{5}$$

which can be interpreted as a local shrinkage estimator pulling the naive estimator towards the simple average of its neighbors.

3 THEORETICAL RESULTS

We will assume that the naive estimators defined by (3) satisfy

$$\max_{i \in [\![B]\!]} \mathrm{MSE}(i, \widehat{\mu}^{\mathrm{NE}}_i) \leq \overline{\sigma}^2. \tag{6}$$

Define the notation

$$G(\tau) := \big\{(i,j) \in [\![B]\!]^2 : \Delta^2_{ij} \leq \tau \overline{\sigma}^2\big\}; \qquad \overline{G}(\tau) := \big\{(i,j) \in [\![B]\!]^2 : \Delta^2_{ij} \geq \tau \overline{\sigma}^2\big\},$$

and two following events:

$$A(\tau) := \Big\{ \max_{(i,j) \in \overline{G}(\tau)} T_{ij} = 1 \Big\}; \qquad B(\tau') := \Big\{ \min_{(i,j) \in G(\tau')} T_{ij} = 0 \Big\};$$

so $\mathbb{P}[A(\tau)]$ is the collective false positive rate of the tests (or family-wise error rate) while $\mathbb{P}[B(\tau')]$ is the collective false negative rate to detect $\Delta^2_{ij} \leq \tau' \overline{\sigma}^2$ (family-wise Type II error rate).

3.1 A General Result under Independence of Estimators and Tests

We start with a result assuming that the tests $(T_{ij})_{(i,j)\in[\![B]\!]^2}$ and the estimators $(\widehat{\mu}_i^{\mathrm{NE}})_{i\in[\![B]\!]}$ are independent. This can be achieved for instance by splitting the original bags into two.

Theorem 3.1. *Assume model* (1) *holds as well as* (2)*, and that* (6) *holds. Furthermore, assume that there exists a family of tests* $(T_{ij})_{(i,j)\in[\![B]\!]^2}$ *that is independent of* $(X_\bullet^{(i)})_{i\in[\![B]\!]}$*. For a fixed constant* $\tau > 0$*, consider the family of estimators* $(\widetilde{\mu}_i)_{i\in[\![B]\!]}$ *defined by* (5) *with respective parameters*

$$\gamma_i := \frac{\tau|V_i^*|}{(1+\tau)|V_i^*|+1}. \tag{7}$$

Then, conditionally to the event $A^c(\tau)$*, it holds*

$$\forall i \in [\![B]\!] : \mathrm{MSE}(i,\widetilde{\mu}_i) \leq \left(\frac{\tau|V_i^*|+1}{(1+\tau)|V_i^*|+1}\right)\overline{\sigma}^2. \tag{8}$$

Let $\mathcal{N}$ *denote the covering number of the set of means* $\{\mu_j, j \in [\![B]\!]\}$ *by balls of radius* $\sqrt{\tau'}\overline{\sigma}/2$*. Then, conditionally to the events* $A^c(\tau)$ *and* $B^c(\tau')$ *(for* $\tau' < \tau$*), it holds*

$$\frac{1}{B}\sum_{i=1}^{B}\mathrm{MSE}(i,\widetilde{\mu}_b) \leq \left(\frac{\tau}{\tau+1} + \frac{\mathcal{N}}{B}\frac{1}{(\tau+1)}\right)\overline{\sigma}^2. \tag{9}$$

The proof can be found in the supplementary material. In a nutshell, conditional to the favorable event $A^c(\tau)$, and because the tests are independent of the estimators, we can use the argument leading to (4), extended to take into account the shrinkage factor γ, and optimize the value of γ to obtain (7), (8). If $B^c(\tau')$ is satisfied as well, we can deduce (9) directly from (8).

Discussion.

- The factor in the individual MSE bound (8) is strictly less than 1 as soon as $|V_i| > 1$. As the number of neighbors $|V_i|$ grows, the factor is larger than but approaches $\tau/(1+\tau)$. Therefore, there is a general trade-off between τ and the number of neighbors in a neighborhood of radius $\sqrt{\tau}\overline{\sigma}$. Nevertheless, in order to aim at possibly significant improvement over naive estimation, a small value of τ should be taken.
- The factor in the averaged MSE bound (9) is also always smaller than 1 (as expected from the individual MSE bound). It has a nice interpretation in terms of the ratio $\mathcal{N}/B$: if $\mathcal{N} \ll B$, the improvement factor will be very close to $\tau/(1+\tau)$. Thus, we collectively can improve over the naive estimation wrt MSE as soon as the set of means has a small covering number (at scale $\sqrt{\tau'}\overline{\sigma}/2$) in comparison to its cardinality. This condition can be met in different structural low complexity situations, e.g. clustered means, means being sparse vectors, set of means on a low-dimensional manifold. Note that the method does not need information about said structure in advance and is in this sense adaptive to it.

3.2 Using the Same Data for Tests and Estimation

We now present a general result in the case where the estimators and tests are not assumed to be independent (e.g. computed from the same data.) To this end we introduce the following additional events:

$$C(\tau) : \left\{\max_{i\neq j}|\langle\widehat{\mu}_i^{\mathrm{NE}} - \mu_i, \widehat{\mu}_j^{\mathrm{NE}} - \mu_j\rangle| > \tau\overline{\sigma}^2\right\}; \qquad C'(\tau) : \left\{\max_i\|\widehat{\mu}_i^{\mathrm{NE}} - \mu_i\|^2 > \overline{\sigma}^2 + \tau\overline{\sigma}^2\right\}.$$

Theorem 3.2. *Assume that there exists a family of tests $(T_{ij})_{(i,j)\in[\![B]\!]^2}$. For a given $\tau > 0$ consider the family of estimators $(\widetilde{\mu}_i)_{i\in[\![B]\!]}$ defined by (5) with respective parameters*

$$\gamma_i := \frac{\tau}{1+\tau}. \tag{10}$$

Then, for $\tau' \geq \tau$, with probability greater than $1 - \mathbb{P}[A(\tau) \cup B(\tau') \cup C(\tau) \cup C'(\tau)]$, it holds

$$\forall i \in [\![B]\!] : \|\widetilde{\mu}_i - \mu_i\|^2 \leq 2\overline{\sigma}^2\bigg(\tau + \frac{\tau + |V_i|^{-1}}{1+\tau}\bigg). \tag{11}$$

Let $\mathcal{N}$ denote the covering number of the set of means $\{\mu_b, b \in [\![B]\!]\}$ by balls of radius $\sqrt{\tau'}\overline{\sigma}/2$. Then, with the same probability as above, it holds

$$\frac{1}{B}\sum_{i=1}^{B}\|\widetilde{\mu}_i - \mu_i\|^2 \leq 2\overline{\sigma}^2\bigg(\tau + \frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{1+\tau}\bigg). \tag{12}$$

The interpretation of the above result is similar to that of Theorem 3.1, with the caveat that the factor in the MSE bound is not always bounded by 1 as earlier; but the qualitative behaviour when τ is small, which is the relevant regime, is the same as previously described.

3.3 The Gaussian Setting

In view of the previous results, the crucial point is whether there exists a family of tests such that the events $A(\tau), B(\tau'), C(\tau), C'(\tau)$ have small probability, for a value of τ significantly smaller than 1, and τ' of the same order as τ (up to an absolute numerical constant). This is what we establish now in the Gaussian setting.

Proposition 3.3. *Assume* (GI) *is satisfied. For a fixed $\alpha \in (0,1)$, define the tests*

$$T_{ij} = \mathbf{1}\Big\{\big\|\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}\big\|^2 \leq \zeta d/N\Big\}, \tag{13}$$

with $\zeta := \Big(\sqrt{2+\tau} - 4\sqrt{\delta}\Big)^2$, where we put $\delta := (2\log B + \log\alpha^{-1})/d$.

Then, provided $\tau \geq \max(C\delta, \sqrt{C\delta})$ (with $C = 10^3$), it holds $\mathbb{P}[A(\tau)] \leq \alpha$, $\mathbb{P}[B(\tau')] \leq \alpha$ with $\tau' = \tau/3$, $\mathbb{P}[C(\tau)] \leq 2\alpha$ and $\mathbb{P}[C'(\tau)] \leq \alpha$.

The above result is significant in combination with Theorems 3.1 and 3.2 when δ is small, which is the case if $\log(B)/d$ is small. The message is the following: in a high-dimensional setting, provided $B \ll e^d$, we can reach a large improvement compared to the naive estimators, if the set of means exhibits structure, as witnessed by a small covering number at scale $d^{\frac{1}{4}}\sqrt{(\log B)/N}$. The best-case scenario is when all the means are tightly clustered around a few values, so that $\mathcal{N}$ is small but B is large, then the improvement in the MSE is by a factor of order $\sqrt{(\log B)/d}$.

3.4 Methodology and Theory in the Kernel Mean Embedding Framework

We recall that the principle of KME posits a reproducing kernel k on an input space $\mathcal{Z}$, corresponding to a feature mapping $\Phi : \mathcal{Z} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, with $k(z,z') = \langle\phi(z), \phi(z')\rangle$. The feature mapping ϕ can be extended to *probability distributions* $\mathbb{P}$ on $\mathcal{Z}$, via $\phi(\mathbb{P}) := \mathbb{E}_{Z\sim\mathbb{P}}[\phi(Z)]$, provided this expectation exists, which can be guaranteed for instance if

ϕ is bounded. This gives rise to an extended kernel on probability distributions via $k(\mathbb{P}, \mathbb{Q}) := \langle \phi(\mathbb{P}), \phi(\mathbb{Q}) \rangle = \mathbb{E}_{(Z,Z') \sim \mathbb{P} \otimes \mathbb{Q}}[k(Z, Z')]$.

As explained in the introduction, if we have a large number of distributions $(\mathbb{P}_i)_{i \in [\![B]\!]}$ for each of which an independent bag $(Z_k^{(i)})_{1 \leq k \leq N_i}$ is available, and we wish to collectively estimate their KMEs, this is an instance of the model (1)-(2) under the transformation $X_k^{(i)} := \phi(Z_k^{(i)})$. The distributions $\mathbb{P}_i$ are replaced by their image distribution through ϕ s.t. $\mu_i = \phi(\mathbb{P}_i)$ and the naive estimators are $\widehat{\mu}_i^{\mathrm{NE}} = \phi(\widehat{\mathbb{P}}_i)$, where $\widehat{\mathbb{P}}_i$ is the empirical measure associated to bag $Z_\bullet^{(i)}$. We will make the assumption that the kernel is bounded, $\sup_{z \in \mathcal{Z}} k(z, z) = \sum_{z \in \mathcal{Z}} \|\phi(z)\|^2 \leq L^2$, resulting in the following "bounded setting":

$$\forall i \in [\![B]\!] : N_i = N \text{ and } \left\|X_k^{(i)}\right\| \leq L, \mathbb{P}_i - \text{a.s.}, k \in [\![N]\!]. \tag{BS}$$

(note in particular that we still assume that all bags have the same size for the theoretical results.)

As always for kernel-based methods, elements of the Hilbert space $\mathcal{H}$ are an abstraction which are never explicitly represented in practice; instead, norms and scalar products between elements, that can be written as linear combinations of sample points, can be computed by straightforward formulas using the kernel. In this perspective, a central object is the *inter-task Gram matrix* K defined as $K_{ij} := k(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_i, \mu_j \rangle, (i, j) \in [\![B]\!]^2$. In the framework of *inference on distributions*, the distributions $\mathbb{P}_i$ act as (latent) training points and the matrix K as the usual kernel Gram matrix for kernel inference. In contrast to what is assumed in standard kernel inference, K is not directly observed but approximated by $\widehat{K}$ s.t. $\widehat{K}_{ij} := \langle \widehat{\mu}_i, \widehat{\mu}_j \rangle$, for some estimators $(\widehat{\mu}_i)_{i \in [\![B]\!]}$ of the true KMEs. The following elementary proposition links the quality of approximation of the means with the corresponding inter-task Gram matrix:

Proposition 3.4. *Assume the model* (1)-(2) *under the assumption* $\left\|X_k^{(i)}\right\| \leq L$ *for all* k, i. *Let* $\widehat{\mu}_i$ *be estimators of* μ_i *bounded by* L, *and the matrices* K *and* $\widehat{K}$ *defined as the Gram matrices of* $(\mu_i)_{i \in [\![B]\!]}$ *and* $(\widehat{\mu}_i)_{i \in [\![B]\!]}$, *respectively. Then*

$$\left\| \frac{1}{B}(K - \widehat{K}) \right\|_{\mathrm{Fr.}}^2 \leq \frac{4L^2}{B} \sum_{i \in [\![B]\!]} \|\mu_i - \widehat{\mu}_i\|^2, \tag{14}$$

where $\|K\|_{\mathrm{Fr.}} := \mathrm{Tr}(KK^T)^{\frac{1}{2}}$ *is the Frobenius norm.*

This result further illustrates the interest of improving the task-averaged squared error.

In order to apply our general results Theorems 3.1 and 3.2, we must again find suitable values of τ (as small as possible) and τ' (as close to τ as possible) so that the probability of the events $A(\tau), B(\tau'), C(\tau), C'(\tau)$ is small, in the setting (BS). In that context, the role of the dimension d will be played by the *effective dimension* $\mathrm{Tr}\,\Sigma / \|\Sigma\|_{op}$, where Σ is the covariance operator for the variable X. More precisely, since this quantity can change from one source distribution to the the other, we will make the following assumption: there exists $d_{\mathrm{eff}} > 0$ such that

$$\forall i \in [\![B]\!] : \qquad d_{\mathrm{eff}} \|\Sigma_i\|_{\mathrm{op}} \leq \mathrm{Tr}\,\Sigma_i \leq N\overline{\sigma}^2. \tag{15}$$

Observe that in view of (3), the upper bound above is merely a reformulation of (6) and, therefore, not a new assumption; the lower bound is.

We consider tests based on the unbiased estimate of the maximum mean discrepancy (MMD; note that the MMD between tasks i and j is exactly Δ_{ij}^2):

$$U_{ij} = \frac{1}{N(N-1)} \sum_{\substack{k,\ell=1 \\ k \neq \ell}}^{N} \left(\langle X_k^{(i)}, X_\ell^{(i)} \rangle + \langle X_k^{(j)}, X_\ell^{(j)} \rangle \right) - \frac{2}{N^2} \sum_{k,\ell=1}^{N} \langle X_k^{(i)}, X_\ell^{(j)} \rangle.$$

Proposition 3.5. *Consider model* (1)*, the bounded setting* (BS) *and assume* (15) *holds. Define*

$$r(t) := 5\left(\sqrt{\left(\frac{1}{d_{\text{eff}}} + \frac{L}{N\overline{\sigma}}\right)t} + \frac{Lt}{N\overline{\sigma}}\right), \tag{16}$$

and

$$\tau_{\min}(t) := r(t)\max\left(\sqrt{2}, r(t)\right). \tag{17}$$

For a fixed $t \geq 1$, define the tests T_{ij} for i, j in $[\![B]\!]^2$

$$T_{ij} := \mathbf{1}\{U_{ij} < \tau\overline{\sigma}^2/2\}. \tag{18}$$

Then, provided $\tau \geq 144\tau_{min}(t)$, it holds

$$\mathbb{P}[A(\tau) \cup B(\tau/4) \cup C(\tau/7) \cup C'(\tau/48)] \leq 14B^2e^{-t}\,.$$

The quantity $r(t)$ above (taking $t = \log(14B^2\alpha^{-1})$, where $1-\alpha$ is the target probability) plays a role analogous to δ in the Gaussian setting (Proposition 3.3). As the bag size N becomes sufficiently large, we expect $\overline{\sigma} = \mathcal{O}(N^{-\frac{1}{2}})$ and, therefore, $\overline{\sigma}N = \mathcal{O}(N^{\frac{1}{2}})$. Hence, provided N is large enough, the quantity $r(t)$ is mainly of the order $\sqrt{\log(B)/d_{\text{eff}}}$. Like in the Gaussian case, this factor determines the potential improvement with respect to the naive estimator, which can be very significant if the effective data dimensionality d_{eff} is large.

From a technical point of view, capturing precisely the role of the effective dimension required us to establish concentration inequalities for deviations of sums of bounded vector-valued variables improving over the classical vectorial Bernstein's inequality of Pinelis and Sakhanenko (1986). We believe this result (see Corollary F.3 in the supplemental) to be of interest of its own and to have potential other applications.

4 EXPERIMENTS AND EVALUATION

We validate our theoretical results in the KME setting[1] on both synthetic as well as real world data. The neighboring kernel means are determined from the tests as described in Eq. (18). More specifically, in practice we use the modification that (i) we adapt the formula for possibly unequal bag sizes, and (ii) in each test T_{ij} we replace $\overline{\sigma}^2$ by the task-dependent unbiased estimate

$$\widehat{\text{MSE}}(i, \widehat{\mu}_i^{\text{NE}}) := \frac{1}{2N_i^2(N_i-1)} \cdot \sum_{k\neq\ell}^{N_i} k(Z_k^{(i)}, Z_k^{(i)}) - 2k(Z_k^{(i)}, Z_\ell^{(i)}) + k(Z_\ell^{(i)}, Z_\ell^{(i)}). \tag{19}$$

We analyze three different variations of our method which we call similarity test based (STB) approaches. `STB-0` corresponds to Eq. (5) with $\gamma = 0$. `STB weight` uses model optimization to find a suitable value for γ, whereas `STB theory` sets γ as defined in Eq. (7). However, here we replaced τ with $c \cdot \tau$, where $c > 0$ is a multiplicative constant, to allow for more flexibility.

We compare their performances to the naive estimation, `NE`, and the regularized shrinkage estimator, `R-KMSE`, (Muandet et al., 2016) which also estimates the KME of each bag separately but shrinks it towards zero. Furthermore, we modified the multi-task averaging approach presented in Feldman et al. (2014) such that it can be used for the estimation of kernel mean embeddings. Similar to our idea, this method shrinks the estimation towards related tasks. However, they require the task similarity to be known. Therefore, we test two options: `MTA const` assumes

[1] In the Gaussian setting, we report numerical results in the Appendix H.

constant similarity for each bag; `MTA stb` uses the proposed test from Eq. (18) to assess the bags for their similarity. See Appendix G for a detailed description of the tested methods.

In the presented results, each considered method has up to two tuning parameters that, in our experiments, are picked in order to optimize averaged test error. Therefore, the reported results can be understood as close to "oracle" performance – the best potential of each method when parameters are close to optimal tuning. While this can be considered unrealistic for practice, a closely related situation can occur in the setting where the user wishes to use the method on test bags of size N, and has at hand a limited number of training bags of much larger size $N' \gg N$. From each such training bag, one can subsample N points, use the method for estimation of the means of all bags of size N (incl. subsampled bags), and monitor the error with respect to the means of the full training bags (of size N', used as a ground truth proxy). This allows a reasonable calibration of the tuning parameters.

4.1 Synthetic Data

The toy data consists of multiple, two-dimensional Gaussian distributed bags $Z_\bullet^{(i)}$ with fixed means but randomly rotated covariance matrices, i.e.

$$Z_\bullet^{(i)} \sim \mathcal{N}\Big(\mathbf{0}, R(\theta_i)\Sigma R(\theta_i)^T\Big) = \mathbb{P}_i \ , \qquad \theta_i \sim \mathcal{U}(-\pi/4, \pi/4),$$

where the covariance matrix $\Sigma = \mathrm{diag}(1, 10)$ is rotated using rotation matrix $R(\theta_i)$ according to angle θ_i. The different estimators are evaluated using the unbiased, squared MMD between the estimation $\widetilde{\mu}_i$ and μ_i as loss. Since μ_i is unknown, it must be approximated by another (naive) estimation $\widehat{\mu}_i^{\mathrm{NE}}(Y_\bullet^{(i)})$ based on independent test bags $Y_\bullet^{(i)}$ from the same distribution as $Z_\bullet^{(i)}$, with $|Y_\bullet^{(i)}| = 1000$. The test bag $Y_\bullet^{(i)}$ has much larger size than the training bag $Z_\bullet^{(i)}$, as a consequence the estimator $\widehat{\mu}_i^{\mathrm{NE}}(Y_\bullet^{(i)})$ has a lower MSE than all considered estimators based on $Z_\bullet^{(i)}$, and can be used as a proxy for the true μ_i.[2] In order to guarantee comparability, all methods use a Gaussian RBF with the kernel width fixed to the average feature-wise standard deviation of the data. Optimal values for the model parameter, e.g. ζ and γ for `STB weight`, are selected such that they minimize the estimation error averaged over 100 trials. Once the values for the parameters are fixed, another 200 trials of data are generated to estimate the final generalization error. Different experimental setups were tested:

(a) **Different Bag Sizes** $B = 50$ and $N_i \in [10, 300]$ for all $i \in [\![B]\!]$,
(b) **Different Number of Bags** $B \in [10, 300]$ and $N_i = 50$ for all $i \in [\![B]\!]$,
(c) **Imbalanced Bags** $B = 50$ and $N_1 = 10, \ldots, N_{50} = 300$,
(d) **Clustered Bags** $N_i, B = 50$ for all $i \in [\![B]\!]$ but the Gaussian distributions are no longer centered around $\mathbf{0}$. Instead, each ten bags form a cluster with the cluster centers equally spaced on a circle. The radius of the circle is varied between 0 and 5, to model different degrees of overlap between clusters.

The results for the experiments on the synthetic data can be found in Figure 1(a) to (d). The estimation of the KME becomes more accurate as the bag size per bag increases. Nevertheless, all of the tested methods provide an increase in estimation performance over the naive estimation, although, the improvement for larger bag sizes decreases for `R-KMSE` and `MTA const`. As expected, methods that use the local neighborhood of the KME yield lower estimation error when the number of available bags increases. Interestingly, this decrease seems to converge towards a capping value, which might reflect the intrinsic dimensionality of the data as indicated by Theorems 3.1 and 3.2

[2] Additionally, the estimation of the squared loss is unbiased if the diagonal entries of the Gram matrix will be included for $Z_\bullet^{(i)}$ but excluded for $Y_\bullet^{(i)}$.

combined with Proposition 3.5. Although we assumed equal bag sizes in the theoretical results, the proposed approaches provide accurate estimations also for the imbalanced setting. Figure 1(c) shows that the improvement is most significant for bags with few samples, which is consistent with results on other multi-task learning problems (see e.g. Feldman et al., 2014). However, when the KME of a bag with many samples is shrunk towards a neighbor with few samples, the estimation can be deteriorated (compare results on (a) with those on (c) for large bag sizes). A similar effect can be seen in the results on the clustered setting. When the bags overlap, a bag from a different cluster might be considered as neighbor which leads to a stronger estimation bias. When the tasks have similar centers or are strictly separated, the methods show similar performance to what is shown in Figure 1(b).

To summarize, `NE` and `R-KMSE` give worst performances because they estimate the kernel means separately. Even though `MTA const` assumes all tasks to be related, it improves the estimation performance even when the bags are not similar. However, the methods that derive the task similarity from the local neighborhood achieve most accurate KME estimations in all of the tested scenarios, especially `STB weight` and `STB theory`.

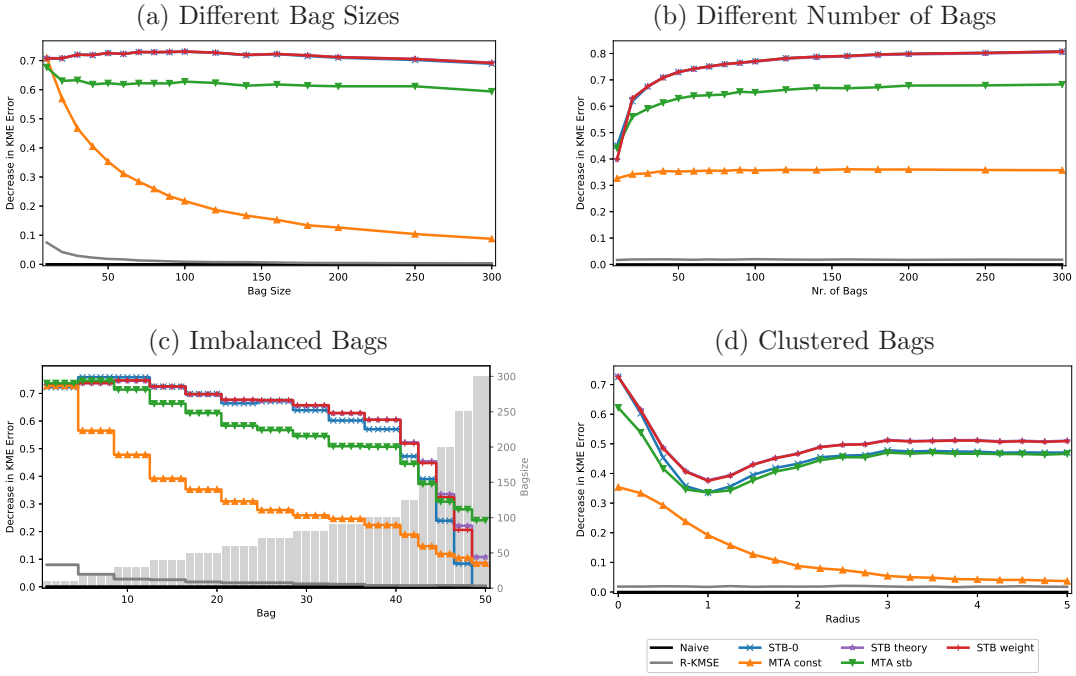

Figure 1: Decrease in KME estimation error compared to `NE` in percent on experimental setups (a) to (d). Higher is better. `STB-0`, `STB weight` and `STB theory` give similar results so that their results might be printed on top of each other.

4.2 Real World Data

We test our methods on a remote sensing data set. The AOD-MISR1 data set is a collection of 800 bags with each 100 samples. The samples correspond to randomly selected pixels from a MISR satellite, where each instance is formed by 12 reflectances from three MISR cameras.[3] It can be

[3]We only use 12 out of 16 features because the remaining four are constant per bag.

used to predict the aerosol optical depth (AOD) which poses an important problem in climate research (Wang et al., 2011).

The data is standardized such that each of the features has unit standard deviation and is centered around zero. In each out of the 100 trials, we randomly subsample 20 samples from each bag, on which the KME estimation is based. This estimation is then compared to the naive estimation on the complete bag. Cross-validation, with 400 bags for training and testing, is used to optimize for the model parameters of each approach and then estimate its error. Again, all methods use a Gaussian RBF with the kernel width fixed to one. The results are shown in Table 1.

Table 1: Decrease in KME estimation error compared to `NE` in percent on the AOD-MISR1 data.

METHOD	%	**METHOD**	%	**METHOD**	%
`R-KMSE`	8.83	`MTA const`	13.92	`STB theory`	21.83
`STB-0`	1.43	`MTA stb`	17.17	`STB weight`	22.73

Again, all of the methods provide a more accurate estimation of the KME than the naive approach. The estimations given by `STB-0` are similar to those of `NE`, because `STB-0` considers very few bags as neighbors. This lets us conclude that the bags are rather isolated than overlapping. `MTA stb`, `STB weight` and `STB theory` might give better estimations because they allow for more flexible shrinkage. Again, `STB weight` and `STB theory` are outperforming the remaining methods.

5 CONCLUSION

In this paper we proposed an improved estimator for the multi-task averaging problem. The estimation is improved by shrinking the naive estimation towards the average of its neighboring means. The neighbors of a task are found by multiple testing so that task similarities must not be known a priori. Provided that appropriate tests exist, we proved that the introduced shrinkage approach yields a lower mean squared error for each task individually and also on average. We show that there exists a family of statistical tests suitable for isotropic Gaussian distributed data or for means that lie in a reproducing kernel Hilbert space. Theoretical analysis shows that this improvement can be especially significant when the (effective) dimension of the data is large, using the property that the typical detection radius of the tests is much better than the standard estimation error in high dimension. This property is particularly important for the estimation of multiple kernel mean embeddings (KME) which is an interesting application relevant for the statistical and machine learning community. The proposed estimator and the theoretical results can naturally be translated to the KME framework.

We tested different variations of the presented approach on synthetic and real world data and compared its performance to other state-of-the-art methods. In all of the conducted experiments, the proposed shrinkage estimators yield the most accurate estimations.

Since the estimation of a KME is often only an intermediate step for solving a final task, as for example in distributional regression (Szabó et al., 2016), further effort must be made to assess whether the improved estimation of the KME also leads to a better final prediction performance. Furthermore, the results on the imbalanced toy data sets have shown that the shrinkage estimator particularly improves the estimation of small bags. However, when the KME of a bag with many samples is shrunk towards a neighbor with low bag size, its estimation might be distorted. There-

fore, another direction for future work will be the development of a similarity test or a weighting scheme that take the bag size into account in a principled way. From a theoretical perspective, we also will investigate if the improvement factor with respect to the naive estimates is optimal in a suitable minimax sense, and if the logarithmic factor $\log(B)$ and the number of tasks appearing in this factor can be removed or alleviated in certain circumstances.

Acknowledgements

The research of HM was funded by the German Ministry for Education and Research as BIFOLD (01IS18025A and 01IS18037A). The research of GB has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901, and by the Agence Nationale de la Recherche (ANR, Chaire IA "BiSCottE"). GB acknowledges various inspiring and motivating discussions with A. Carpentier, U. Dogan, C. Giraud, V. Koltchinskii, G. Lugosi, A. Maurer, G. Obozinski, C. Scott.

A Proof of Theorem 3.1

We argue conditional to the tests, below expectations are taken with respect to the samples $(X_\bullet^{(b)})_{b\in[\![B]\!]}$ only. Assume the event $A^c(\tau)$ holds, implying for all i:

$$j \in V_i \Rightarrow \Delta_{ij}^2 \le \tau\sigma^2. \tag{20}$$

Take $i = 1$ without loss of generality, and denote $V = V_1, V^* = V_1 \setminus \{1\}$, and $v = |V_1|$. We also put $\eta = 1 - \gamma$. We use an argument similar to that leading to (4) using independence of the bags, triangle inequality and (20):

$$\begin{aligned}
\mathrm{MSE}(1,\widetilde{\mu}_1) &= \mathbb{E}\left[\left\|(1-\eta)(\widehat{\mu}_1^{\mathrm{NE}} - \mu_1) + \frac{\eta}{v}\sum_{j\in V}(\widehat{\mu}_i^{\mathrm{NE}} - \mu_1)\right\|^2\right] \\
&= \frac{\eta^2}{v^2}\left(\left\|\sum_{i\in V^*}(\mu_i - \mu_1)\right\|^2 + \sum_{i\in V^*}\mathbb{E}\left[\left\|\mu_i - \widehat{\mu}_i^{\mathrm{NE}}\right\|^2\right]\right) \\
&\qquad + (1-\eta(1-v^{-1}))^2\mathbb{E}\left[\left\|\widehat{\mu}_1^{\mathrm{NE}} - \mu_1\right\|^2\right] \\
&\le \sigma^2\Big(\frac{\eta^2}{v^2}\big((v-1)^2\tau + (v-1)\big) + (1-\eta(1-v^{-1}))^2\Big) \\
&= \sigma^2\big(\eta^2(1-v^{-1})\big((1-v^{-1})\tau + 1\big) - 2\eta(1-v^{-1}) + 1\big).
\end{aligned}$$

The optimal value of $\gamma = 1 - \eta$ is given by (7) and gives rise to (8).

Assume additionally that $B^c(\tau')$ holds. Let $\varepsilon := \sqrt{\tau'}\sigma/2$ and let $\mathcal{C} := \{x_1, \dots, x_{\mathcal{N}}\}$ be an ε-covering of the set of means. Let $\pi(i)$ be the index of the element of $\mathcal{C}$ closest to μ_i, and $N_k := \{b \in [\![B]\!] : \pi(b) = k\}$, $i \in [\![\mathcal{N}]\!]$. By the triangle inequality, for any $i \in [\![\mathcal{N}]\!]$, $b \in N_i$ one has $|V_b| \ge \big|N_{\pi(i)}\big|$. Hence averaging (8) over i we get

$$\begin{aligned}
\frac{1}{B}\sum_{b=1}^{B}\mathrm{MSE}(b,\widetilde{\mu}_b) &\le \frac{\sigma^2}{B}\sum_{i\in[\![B]\!]}\frac{\tau(|N_{\pi(i)}|-1)+1}{1+(1+\tau)(|N_{\pi(i)}|-1)} \\
&= \frac{\sigma^2}{B}\sum_{k\in[\![\mathcal{N}]\!]}\frac{|N_k|(\tau(|N_k|-1)+1)}{1+(1+\tau)(|N_k|-1)}.
\end{aligned}$$

The above take the form $\sum_k f(|N_k|)$, and it is straightforward to check that f is convex. Since it holds $1 \le |N_k| \le B - \mathcal{N} + 1$ for all k, and $\sum_{k\in[\![\mathcal{N}]\!]}|N_k| = B$, the maximum of the above expression is attained for an extremal point of this convex domain, i.e., by symmetry, $N_1 = B - \mathcal{N} + 1$ and $N_k = 1$ for $k \ge 2$. Therefore

$$\begin{aligned}
\frac{1}{B}\sum_{b=1}^{B}\mathrm{MSE}(b,\widetilde{\mu}_b) &\le \frac{\sigma^2}{B}\left((\mathcal{N}-1) + \frac{(B-\mathcal{N}+1)((B-\mathcal{N})\tau+1)}{(B-\mathcal{N})(1+\tau)+1}\right) \\
&= \frac{\sigma^2}{B}\left(\mathcal{N} + \frac{(B-\mathcal{N})^2\tau}{(B-\mathcal{N})(1+\tau)+1}\right) \\
&\le \sigma^2\left(\frac{\tau}{\tau+1} + \frac{\mathcal{N}}{B}\frac{1}{\tau+1}\right).
\end{aligned}$$

□

B Proof of Theorem 3.2

We follow the same general line as in theorem 3.1. Assume the event $A^c(\tau) \cap B^c(\tau') \cap C^c(\tau) \cap C'^c(\tau)$ holds. Take $i = 1$ without loss of generality, and denote $V = V_1, V^* = V_1 \setminus \{1\}$, and $v = |V_1|$. We still put $\eta = 1 - \gamma$. Then

$$\begin{aligned}\|\widetilde{\mu}_1 - \mu_1\|^2 &= \left\|(1-\eta)(\widehat{\mu}_1^{\mathrm{NE}} - \mu_1) + \frac{\eta}{v}\sum_{j\in V}(\widehat{\mu}_j^{\mathrm{NE}} - \mu_1)\right\|^2 \\ &\leq 2\left(\left\|(1-\eta(1-v^{-1}))(\widehat{\mu}_1^{\mathrm{NE}} - \mu_1) + \frac{\eta}{v}\sum_{j\in V^*}(\widehat{\mu}_j^{\mathrm{NE}} - \mu_j)\right\|^2 + \frac{\eta^2}{v^2}\left\|\sum_{j\in V^*}\mu_j - \mu_1\right\|^2\right).\end{aligned}$$

Let us upper bound the different terms. Because $j \in V$, we know that $\Delta_{j1} \leq \tau\overline{\sigma}^2$, so by the triangle inequality

$$\frac{\eta}{v}\left\|\sum_{j\in V^*}\mu_j - \mu_1\right\| \leq \frac{\eta}{v}\sum_{j\in V^*}\|\Delta_{ij}\| \leq \eta(1-v^{-1})\sqrt{\tau}\overline{\sigma}.$$

Let us develop the other term :

$$\begin{aligned}&\left\|(1-\eta(1-v^{-1}))(\widehat{\mu}_1^{\mathrm{NE}} - \mu_1) + \frac{\eta}{v}\sum_{j\in V^*}(\widehat{\mu}_j^{\mathrm{NE}} - \mu_j)\right\|^2 \\ &\quad= (1-\eta(1-v^{-1}))^2\|\widehat{\mu}_1^{\mathrm{NE}} - \mu_1\|^2 + \frac{2\eta(1-\eta(1-v^{-1}))}{v}\sum_{j\in V^*}\langle\widehat{\mu}_1^{\mathrm{NE}} - \mu_1, \widehat{\mu}_j^{\mathrm{NE}} - \mu_j\rangle \\ &\qquad+ \frac{\eta^2}{v^2}\sum_{j\neq k\in V^*}\langle\widehat{\mu}_j^{\mathrm{NE}} - \mu_j, \widehat{\mu}_k^{\mathrm{NE}} - \mu_k\rangle + \frac{\eta^2}{v^2}\sum_{j\in V^*}\|\widehat{\mu}_j^{\mathrm{NE}} - \mu_j\|^2 \\ &\quad\leq \overline{\sigma}^2\Big[(1-\eta(1-v^{-1}))^2(1+\tau) + 2\eta(1-\eta(1-v^{-1}))(1-v^{-1})\tau \\ &\qquad\qquad+ \eta^2(1-v^{-1})^2\tau + \eta^2 v^{-1}(1-v^{-1})(1+\tau)\Big].\end{aligned}$$

Let us associate the two expressions, we obtain that :

$$\|\widetilde{\mu}_1 - \mu_1\|^2 \leq 2\overline{\sigma}^2\Big[\tau + 1 - 2(1-v^{-1})\eta + (1-v^{-1})(1+\tau)\eta^2\Big].$$

The expression is minimal when $\eta = (1+\tau)^{-1}$. By the same arguments about using covering numbers as in the proof of Theorem 3.1, we obtain that with probability greater than $1 - \mathbb{P}[A(\tau) \cup B(\tau') \cup C(\tau) \cup C'(\tau)]$:

$$\begin{aligned}\frac{1}{B}\sum_{i\in[\![B]\!]}\|\widetilde{\mu}_i - \mu_i\|^2 &\leq \frac{2\overline{\sigma}^2}{B}\sum_{i\in[\![B]\!]}\tau + \frac{\tau + |V_i|^{-1}}{1+\tau} \\ &\leq 2\overline{\sigma}^2\left(\tau + \frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{1+\tau}\right).\end{aligned}$$

□

C Proof of Proposition 3.3

Recall that we assume the (GI) model. We first consider the behavior of a single test $T_{ij} = \mathbf{1}\Big\{\big\|\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}\big\|^2 \leq \zeta\overline{\sigma}^2\Big\}$, where $\overline{\sigma}^2 := d/N$, we also put $\Delta^2 = \Delta_{ij}^2$ for short. The random variable $Z := \widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}$ is distributed as $\mathcal{N}(\mu_i - \mu_j, 2n^{-1}I_d)$ by independence of the bags. From classical concentration results for chi-squared variables recalled as Proposition E.1 in Section E, for any $\alpha \in (0,1)$ either of the inequalities below hold with probability $1-\alpha$:

$$\sqrt{\Delta^2 + 2\overline{\sigma}^2} - 4\overline{\sigma}\sqrt{\frac{\log\alpha^{-1}}{d}} \leq \|Z\| \leq \sqrt{\Delta^2 + 2\overline{\sigma}^2} + 2\overline{\sigma}\sqrt{\frac{\log\alpha^{-1}}{d}}. \tag{21}$$

Put $\delta := (\log\alpha^{-1})/d$ for short.

We start with analyzing Type I error: if $\Delta^2 \geq \tau\overline{\sigma}^2$, then the above lower bound implies $\|Z\|^2 \geq \overline{\sigma}^2\big(\sqrt{2+\tau} - 4\sqrt{\delta}\big)^2$, so $T_{ij} = 0$ if we choose $\zeta := \big(\sqrt{2+\tau} - 4\sqrt{\delta}\big)^2$. By union bound over $(i,j) \in [\![B]\!]^2$, with this choice we guarantee that $\mathbb{P}[A(\tau)] \leq \alpha$ if we replace α by αB^2 (i.e. take $\delta = (2\log B + \log\alpha^{-1})/d$). This establishes the bound on family-wise type I error.

We now analyze type II error: assume now that we have picked $\zeta := \big(\sqrt{2+\tau} - 4\sqrt{\delta}\big)^2$, with $\tau \geq \max(C\delta, \sqrt{C\delta})$, $C = 1000$, and assume $\Delta^2 \leq \tau'\overline{\sigma}^2$. Then assuming the upper bound in (21) is satisfied, we ensure $T_{ij} = 1$ provided

$$\sqrt{\tau' + 2} \leq \sqrt{\tau + 2} - 6\sqrt{\delta}.$$

Note that the condition on τ ensures that the above right-hand-side is positive. Taking squares and further bounding, a sufficient condition for the above is $\tau' \leq \tau - 12\sqrt{(2+\tau)\delta}$. Using the condition on τ, it holds

$$12\sqrt{(2+\tau)\delta} \leq 12\sqrt{3C^{-1}}\tau \leq \frac{2}{3}\tau,$$

hence $\tau' \leq \tau/3$ is a sufficient condition. This ensures, by the union bound, that $\mathbb{P}[B(\tau')] \leq \alpha$ when replacing δ by $\delta' = (2\log B + \log\alpha^{-1})/d$ as above.

We now turn to controlling the probability of the events $C(\tau)$ and $C'(\tau)$. For fixed i,j put $X_1 = \widehat{\mu}_i^{\mathrm{NE}} - \mu_i$, $X_2 = \widehat{\mu}_j^{\mathrm{NE}} - \mu_j$. Under the (GI) model, X_1, X_2 are independent $\mathcal{N}(0, N^{-1}I_d)$. Applying the result of Proposition E.2, we obtain that for $\alpha \in (0,1)$, we have probability at least $1-2\alpha$:

$$|\langle X_i, X_j\rangle| \leq \overline{\sigma}^2\big(\sqrt{2\delta} + \delta\big),$$

where we have put $\delta := (\log\alpha^{-1})/d$ as previously. As soon as $\tau \geq \max(C\delta, \sqrt{C\delta})$, $(C \geq 1)$ we obtain $|\langle X_i, X_j\rangle| \leq 3\tau\overline{\sigma}^2/\sqrt{C}$ on the above event, implying that the event $C(\tau)$ is a fortiori satisfied for $C = 10^3$.

From estimate (24) in Proposition E.1, we have with probability at least $1-\alpha$:

$$\|X_1\| \leq \overline{\sigma}^2\big(1 + \sqrt{2\delta}\big) \leq \overline{\sigma}^2\big(1 + 2\tau C^{-1}\big),$$

under the same condition on τ as above. As previously, by the union bound the above estimates are true simultaneously for all i,j with the indicated probabilities if we replace δ by $\delta' = (2\log B + \log\alpha^{-1})/d$, and $C'(\tau)$ is satisfied when taking $C = 10^3$. □

D Results in the Bounded Setting (for KME Estimation)

D.1 Proof of Proposition 3.4

$$\begin{aligned}
\big\|(K-\widehat{K})\big\|_{\mathrm{Fr.}}^2 &= \sum_{(i,j)\in[\![B]\!]^2} (\langle \mu_i,\mu_j\rangle - \langle \widehat{\mu}_i,\widehat{\mu}_j\rangle)^2 \\
&= \sum_{(i,j)\in[\![B]\!]^2} (\langle \mu_i-\widehat{\mu}_i,\mu_j\rangle + \langle \widehat{\mu}_i,\mu_j-\widehat{\mu}_j\rangle)^2 \\
&\le 2\sum_{(i,j)\in[\![B]\!]^2} (\langle \mu_i-\widehat{\mu}_i,\mu_j\rangle^2 + \langle \widehat{\mu}_i,\mu_j-\widehat{\mu}_j\rangle^2) \\
&\le 2L^2\sum_{(i,j)\in[\![B]\!]^2} (\|\mu_i-\widehat{\mu}_j\|^2 + \|\mu_j-\widehat{\mu}_j\|^2) \\
&\le 4L^2B\sum_{i\in[\![B]\!]} \|\mu_i-\widehat{\mu}_j\|^2.
\end{aligned}$$

□

D.2 Proof of Proposition 3.5

Recall the notation

$$r(t) = 5\left(\sqrt{\left(\frac{1}{d_{\mathrm{eff}}}+\frac{L}{N\overline{\sigma}}\right)t} + \frac{Lt}{N\overline{\sigma}}\right), \tag{22}$$

and

$$\tau_{\min}(t) := r(t)\max\Big(\sqrt{2}, r(t)\Big). \tag{23}$$

Introduce the notation $q(t) := \overline{\sigma} r(t)$; $\xi(t) := \overline{\sigma}^2\tau_{\min}(t) = q(t)\max(\sqrt{2}\overline{\sigma}, q(t))$. Let $i,j \in [\![B]\!]^2$ be fixed and $t \ge 1$. We put $\tau = \lambda^2\tau_{\min}(t)$ with $\lambda \ge 12$.

Suppose that $\|\Delta_{ij}\|^2 > \tau\overline{\sigma}^2 = \lambda^2\tau_{\min}\overline{\sigma}^2 = \lambda^2\xi(t)$. We use the concentration inequality (40) for bounded variables, proved in Section F, and obtain that with probability greater than $1-8e^{-t}$, and using the definition of $\xi(t)$:

$$U_{ij} \ge \|\Delta_{ij}\|^2 - 2\|\Delta_{ij}\|q(t) - 8\sqrt{2\overline{\sigma}^2}q(t) - 32q^2(t) \ge \|\Delta_{ij}\|\Big(\|\Delta_{ij}\| - 2q(t)\Big) - 40\xi(t)\,.$$

(To be more precise, (40) proves the above estimate for the value of $q(t)$ defined by (37), the value of $q(t)$ defined in the present proof is an upper bound for it, so the above also holds.)

Observe $\|\Delta_{ij}\| \ge \lambda\sqrt{\xi(t)} \ge 12\sqrt{\xi(t)} \ge 2q(t)$. By monotonicity in $\|\Delta_{ij}\|$ under that condition, it holds $\|\Delta_{ij}\|\Big(\|\Delta_{ij}\| - 2q(t)\Big) \ge \sqrt{\lambda\xi(t)}(\lambda\sqrt{\xi(t)} - 2q(t)) \ge \lambda(\lambda-2)\xi(t)$. That leads to

$$U_{ij} \ge (\lambda^2 - 2\lambda - 40)\xi(t) \ge (\lambda^2/2)\xi(t) = (\tau/2)\overline{\sigma}^2,$$

where we have used that $\lambda^2 - 2\lambda - 40 \ge \lambda^2/2$ for $\lambda \ge 12$. So

$$\mathbb{P}\big[\|\Delta_{ij}\|^2 > \tau\overline{\sigma}^2 \quad \text{and} \quad T_i = 1\big] \le 8e^{-t}.$$

Suppose now $\|\Delta_{ij}\|^2 < (\tau/4)\overline{\sigma}^2 = (\lambda^2/4)\xi(t)$. Then, according to the concentration ineqality (39), with probability greater than $1 - 8e^{-t}$, it holds

$$\begin{aligned} U_{ij} &\leq \|\Delta_{ij}\|^2 + 2\|\Delta\|q(t) + 2\sqrt{2\overline{\sigma}^2}q(t) + 11q^2(t) \\ &\leq \Big(\lambda^2/4 + \lambda + 13\Big)\xi(t) \\ &\leq (\lambda^2/2)\xi(t) = (\tau/2)\overline{\sigma}^2. \end{aligned}$$

We have used that $\lambda^2/4 + \lambda + 13 \leq \lambda^2/2$ for $\lambda \geq 12$. So

$$\mathbb{P}\big[\|\Delta_{ij}\|^2 < \tau\overline{\sigma}^2/4 \quad \text{and} \quad T_i = 0\big] \leq 2e^{-t}.$$

An union bound over $(i,j) \in [\![B]\!]^2$ gives that

$$\mathbb{P}[A(\tau) \cup B(\tau/4)] \leq 8B^2e^{-t}\,.$$

Remarking that

$$\overline{\sigma}^2\tau/7 \geq 20q(t)\max(q(t), \sqrt{2\overline{\sigma}^2}) \quad \text{and} \quad \overline{\sigma}^2\tau/48 \geq 3q(t)\max(q(t), \sqrt{2\overline{\sigma}^2}) \geq 2q(t)\sqrt{2\overline{\sigma}^2} + q^2(t)$$

and using the concentration inequalities (30) and (36) gives

$$\mathbb{P}[C(\tau/7)] \leq 6(B^2 - B)e^{-t}\,, \quad \text{and} \quad \mathbb{P}[C'(\tau/48)] \leq Be^{-t}\,.$$

□

E Concentration Results in the Gaussian Setting

Proposition E.1. *Let Z be a normal $\mathcal{N}(\mu, \sigma^2 I_d)$ random variable in $\mathbb{R}^d$. Then for any $t \geq 0$:*

$$\mathbb{P}\left[\|Z\| \geq \sqrt{\|\mu\|^2 + \sigma^2 d} + \sigma\sqrt{2t}\right] \leq e^{-t}\,, \tag{24}$$

and

$$\mathbb{P}\left[\|Z\| \leq \sqrt{\|\mu\|^2 + \sigma^2 d} - 2\sigma\sqrt{2t}\right] \leq e^{-t}. \tag{25}$$

Proof. The stated inequalities are direct consequences of classical deviation inequalities for (non-central) χ^2 variables. Put $\lambda := \|\mu\|^2$, then for the upper deviation bound, Lemma 8.1 of (Birgé, 2001) states that

$$\mathbb{P}\Big[\|Z\|^2 \geq \lambda + d\sigma^2 + 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t} + 2\sigma^2 t\Big] \leq e^{-t}\,,$$

and we have

$$\lambda + d\sigma^2 + 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t} + 2\sigma^2 t \leq \Big(\sqrt{\lambda + d\sigma^2} + \sigma\sqrt{2t}\Big)^2\,,$$

implying (24). For the lower deviation bound, Lemma 8.1 of (Birgé, 2001) states that

$$\mathbb{P}\Big[\|Z\|^2 \leq \lambda + d\sigma^2 - 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t}\Big] \leq e^{-t}\,,$$

and we have

$$\Big(\lambda + d\sigma^2 - 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t}\Big)_+ \geq \sqrt{\lambda + d\sigma^2}\Big(\sqrt{\lambda + d\sigma^2} - 2\sigma\sqrt{2t}\Big)_+ \geq \Big(\sqrt{\lambda + d\sigma^2} - 2\sigma\sqrt{2t}\Big)_+^2,$$

leading to (25). □

Proposition E.2. *Let X_1, X_2 be independent $\mathcal{N}(0, \sigma^2 I_d)$ variables in dimension d. Then for any $t \geq 0$:*

$$\mathbb{P}\Big[\langle X_1, X_2\rangle \geq \sigma^2\Big(\sqrt{2dt} + t\Big)\Big] \leq e^{-t}. \tag{26}$$

Proof. Without loss of generality assume $\sigma^2 = 1$. For two independent one-dimensional Gaussian variables G_1, G_2, one has for any $\lambda \in [0, 1]$:

$$\mathbb{E}[\exp \lambda G_1 G_2] = \mathbb{E}[\mathbb{E}[\exp \lambda G_1 G_2 | G_2]] = \mathbb{E}\Big[\exp \frac{\lambda^2}{2} G_2^2\Big] = \frac{1}{\sqrt{1-\lambda^2}}\,,$$

so that

$$\log \mathbb{E}[\exp \lambda \langle X_1, X_2\rangle] = \frac{d}{2}(-\log(1-\lambda^2)) \leq \frac{d}{2}\frac{\lambda^2}{(1-\lambda)}\,.$$

Applying Lemma 8.2 of (Birgé, 2001) gives (26). □

F Concentration Results in the Bounded Setting

Studying concentration in the kernel setting means having concentration results of bounded variables taking values in a separable Hilbert space. Recall that $k(x, y) = \langle \phi(z), \phi(z')\rangle_{\mathcal{H}}$ for all z,z' in $\mathcal{H}$, so that if k is bounded by L^2, then the map ϕ is bounded by L. To obtain concentration results, we will use Talagrand's inequality.

Theorem F.1 (Talagrand's inequality)**.** *Let $X_1^s, ..., X_N^s$ be iid real random variables indexed by $s \in S$ where S is a countable index set, and L be a positive constant such that:*

$$\mathbb{E}[X_k^s] = 0\,, \quad \textit{and} \quad |X_k^s| \leq L \ a.s. \quad \forall k \in [\![N]\!],\, s \in S\,.$$

Let us note $Z = \sup_{s\in S} \sum_{k=1}^N X_k^s$, then for all $t \geq 0$:

$$\mathbb{P}\Big[Z - \mathbb{E}[Z] \geq 2\sqrt{(2v + 16L\mathbb{E}[Z])t} + 2Lt\Big] \leq e^{-t};$$
$$\mathbb{P}\Big[-Z + \mathbb{E}[Z] \geq 2\sqrt{(4v + 32L\mathbb{E}[Z])t} + 4Lt\Big] \leq e^{-t},$$

where $v = \sup_{s\in S} \sum_{k=1}^N \mathbb{E}\big[(X_k^s)^2\big]$.

Talagrand's inequality appeared originally in Talagrand (1996), with the above form (using additional symmetrization and contraction arguments from Ledoux and Talagrand, 1991) appearing in Massart (2000). The constants in the upper deviation bound have been improved by Rio (2002) and Bousquet (2002), however no such improvement is available for lower devations as far as we know. The above version is taken from Massart (2007) p. 169–170, (5.45) and (5.46) combined with (5.47) there.

Because a Hilbertian norm can be viewed as a supremum, we can use Talagrand's inequality to obtain a concentration inequality for the norm of the sum of bounded Hilbert-valued random variables.

Proposition F.2. *Let $(Z_k)_{1\leq k\leq N}$ be i.i.d. random variables taking values in a separable Hilbert space $\mathcal{H}$, whose norm is bounded by L a.s. Let μ and Σ denote their common mean and covariance operator. Let*

$$V = \left\|\frac{1}{N}\sum_{k=1}^N Z_k\right\|, \quad \textit{and} \quad V_c = \left\|\frac{1}{N}\sum_{k=1}^N Z_k - \mu\right\|.$$

Then for any $t \geq 0$:

$$\mathbb{P}\Big[V^2 \geq \|\mu\|^2 + \big(\mathbb{E}[V_c] + q_\Sigma(t)\big)^2 + 2\|\mu\| q_\Sigma(t)\Big] \leq 2e^{-t}\,, \tag{27}$$

and

$$\mathbb{P}\Big[V^2 \leq \|\mu\|^2 + \big(\mathbb{E}[V_c] - 2q_\Sigma(t)\big)_+^2 - 2\|\mu\| q_\Sigma(t)\Big] \leq 2e^{-t}\,, \tag{28}$$

where

$$q_\Sigma(t) = 2\sqrt{\left(\frac{2\|\Sigma\|_{\mathrm{op}}}{N} + 16L\frac{\sqrt{\operatorname{Tr}\Sigma}}{N^{3/2}}\right)t} + \frac{2L}{N}t. \tag{29}$$

Proof. Let us denote $q(t)$ for $q_\Sigma(t)$ for this proof. We start with bounding the deviations of V_c. Observe that

$$V_c = \sup_{\|u\|_{\mathcal{H}}=1} \frac{1}{N}\sum_{k=1}^N \langle u, Z_k - \mu\rangle,$$

where the supremum can be restricted to u in a dense countable subset $\mathcal{S}$ of the unit sphere, since $\mathcal{H}$ is separable. We can therefore apply Talagrand's inequality with $X_k^u := N^{-1}\langle u, Z_k - \mu\rangle$; it holds $|X_k^u| \leq L/N$, and note that since $\Sigma = \mathbb{E}[(Z-\mu)\otimes(Z-\mu)^*]$, it holds

$$\mathbb{E}\big[(X_k^u)^2\big] = N^{-2}\mathbb{E}\Big[\langle u, Z_k - \mu\rangle^2\Big] = N^{-2}\langle u, \Sigma u\rangle,$$

so that $\sup_{u\in\mathcal{S}}\sum_{k=1}^N \mathbb{E}\big[(X_k^u)^2\big] = N^{-1}\|\Sigma\|_{\mathrm{op}}$. Furthermore, $\mathbb{E}[V_c] \leq N^{-\frac{1}{2}}\sqrt{\operatorname{Tr}\Sigma}$ by Jensen's inequality, which we use to further bound the deviation term by $q(t)$.

By Theorem F.1, with probability greater than $1-e^{-t}$ for $t \geq 0$, it holds

$$V_c \leq \mathbb{E}[V_c] + q(t)\,, \tag{30}$$

and with probability greater than $1-e^{-t}$,

$$V_c \geq \mathbb{E}[V_c] - 2q(t)\,. \tag{31}$$

We turn to bounding the deviations of $V^2 - \|\mu\|^2$. Observe

$$V^2 - \|\mu\|^2 = V_c^2 + \frac{2}{N}\sum_{k=1}^N \langle Z_k - \mu, \mu\rangle. \tag{32}$$

Using Bernstein's inequality for the variables $W_i = \langle Z_i - \mu, \mu\rangle$, satifying $\mathbb{E}[W_i] = 0$, $\mathbb{E}\big[W_i^2\big] = \langle \mu, \Sigma\mu\rangle \leq \|\Sigma\|_{\mathrm{op}}\|\mu\|^2$, and $|W_i| \leq L\|\mu\|$, we have that with probability greater than $1-e^{-t}$, for $t \geq 0$:

$$\frac{1}{N}\sum_{i=1}^N \langle Z_i - \mu, \mu\rangle \leq \|\mu\|\left[\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}t}{N}} + \frac{4Lt}{3N}\right] \leq \|\mu\| q(t)\,. \tag{33}$$

Combining inequality (33) with (32) and (30) gives that with probability greater than $1-2e^{-t}$:

$$V^2 - \|\mu\|^2 \leq \big(\mathbb{E}[V_c] + q(t)\big)^2 + 2\|\mu\| q(t)\,,$$

and, combining (33), (32) and (31), we have with probability greater than $1-2e^{-t}$:

$$V^2 - \|\mu\|^2 \geq \big(\mathbb{E}[V_c] - 2q(t)\big)_+^2 - 2\|\mu\| q(t)\,.$$

□

Corollary F.3. *Using the setting and notation of Proposition F.2, we have*

$$-2q_\Sigma(1) + \sqrt{\frac{\operatorname{Tr}\Sigma}{N}} \leq \mathbb{E}[V_c] \leq \sqrt{\frac{\operatorname{Tr}\Sigma}{N}}\,.$$

As a consequence, for any $t > 0$,

$$\mathbb{P}\left[V^2 \geq \|\mu\|^2 + \left(\sqrt{\frac{\operatorname{Tr}\Sigma}{N}} + q_\Sigma(t)\right)^2 + 2\|\mu\| q_\Sigma(t)\right] \leq 2e^{-t}\,, \tag{34}$$

and for any $t \geq 1$,

$$\mathbb{P}\left[V^2 \leq \|\mu\|^2 + \left(\sqrt{\frac{\operatorname{Tr}\Sigma}{N}} - 4q_\Sigma(t)\right)_+^2 - 2\|\mu\| q_\Sigma(t)\right] \leq 2e^{-t}\,, \tag{35}$$

Remark. To the expert reader, we want to point out that the above concentration estimates are sharper than the Bernstein's concentration inequality for vector random variables due to Pinelis and Sakhanenko (1986) (Corollary 1 there) and which has found many uses in the recent literature on kernel methods. The reason is that in Pinelis and Sakhanenko's result, which concerns deviations of the centered process V_c, the deviation term (in factor of t) for V_c is proportional to $\sqrt{\operatorname{Tr}\Sigma/N}$. The inequality of Pinelis and Sakhanenko also only bounds upper deviations.

In contrast, in the above result, the term $\sqrt{\operatorname{Tr}\Sigma/N} = \mathbb{E}\big[\|V_c\|^2\big]^{\frac{1}{2}}$ appears with constant 1, and the main deviation term (in factor of t) only involves $\sqrt{\|\Sigma\|_{\mathrm{op}}/N}$, which is better by a factor of $1/\sqrt{d_{\mathrm{eff}}}$. We also obtain the informative lower deviation bound (35).

To summarize, Pinelis and Sakhanenko (1986)'s inequality controls the upper deviations of V_c from zero in terms of a factor of its expectation, while the above concentration inequalities control the two-sided deviations of V_c^2 from its *expectation*, which is $\operatorname{Tr}\Sigma/N$, in terms of a factor of its typical deviation, which is $\|\Sigma\|_{\mathrm{op}}/N$.

This improvement makes the above bound first-order correct and mimic more closely the Gaussian chi-squared deviation phenomenon of Proposition E.1. This sharpness (and the fact that we get a control for two-sided deviations) is crucial in order to be able to capture the behavior of the effective dimension, see in particular Proposition F.5 below for the analysis of the MMD U-statistic, for which the exact cancellation of the first order terms is paramount.

Proof. The upper bound of the mean of V_c is given directly by Jensen's inequality. For the lower bound, we can rewrite Talagrand's inequality (30) equivalently under the following form: there exists ξ, an exponential random variable of parameter 1, such that almost surely

$$V_c \leq \mathbb{E}[V_c] + q_\Sigma(\xi) = \mathbb{E}[V_c] + \alpha\sqrt{\xi} + \beta\xi,$$

where α and β are given by (29). Taking the square and then the mean gives :

$$\begin{aligned}\mathbb{E}\big[V_c^2\big] &\leq \mathbb{E}\left[\left(\mathbb{E}[V_c] + \alpha\sqrt{\xi} + \beta\xi\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\mathbb{E}[V_c] + (\alpha+\beta)\sqrt{\xi}\right)^2 + 2(\alpha+\beta)\mathbb{E}[V_c]\xi + (\alpha+\beta)^2\xi^2\right].\end{aligned}$$

We can use now the concavity of the function $\xi \mapsto \left(\mathbb{E}[V_c] + (\alpha+\beta)\sqrt{\xi}\right)^2$ and Jensen's inequality, obtaining

$$\mathbb{E}\big[V_c^2\big] \leq \left(\mathbb{E}[V_c] + (\alpha+\beta)\right)^2 + 2(\alpha+\beta)\mathbb{E}[V_c] + 2(\alpha+\beta)^2 \leq \left(\mathbb{E}[V_c] + 2(\alpha+\beta)\right)^2.$$

Because $\mathbb{E}\big[V_c^2\big] = \operatorname{Tr}\Sigma/N$, and $(\alpha+\beta) = q_\Sigma(1)$ by definition, we obtain that

$$\mathbb{E}[V_c] \geq \sqrt{\frac{\operatorname{Tr}\Sigma}{N}} - 2q_\Sigma(1).$$

If $t \geq 1$, it holds $q(t) \geq q(1)$ and we can plug in the above estimates for $\mathbb{E}[V_c]$ into (27) and (28) to obtain (34) and (35), respectively (note that the condition $t \geq 1$ is only needed for the lower devation bound). □

Proposition F.4. *Let $(X_k)_{1\leq k\leq N} \overset{i.i.d.}{\sim} X$ and $(Y_k)_{1\leq k\leq N} \overset{i.i.d.}{\sim} Y$ be independent families of centered random variables bounded by L in a separable Hilbert space $\mathcal{H}$. Let Σ_X and Σ_Y be their respective covariance operators, $\overline{\sigma}^2$ and d_{eff} such that*

$$\max(\operatorname{Tr}\Sigma_X, \operatorname{Tr}\Sigma_Y)/N \leq \overline{\sigma}^2\,;$$
$$\min\left(\frac{\operatorname{Tr}\Sigma_X}{\|\Sigma_X\|_{op}}, \frac{\operatorname{Tr}\Sigma_Y}{\|\Sigma_Y\|_{op}}\right) \geq d_{\text{eff}}.$$

Then for any $t \geq 0$:

$$\mathbb{P}\left[\left\langle \frac{1}{N}\sum_{k=1}^N X_k, \frac{1}{N}\sum_{k=1}^N Y_k \right\rangle \geq 20q(t)\max(\overline{\sigma}, q(t))\right] \leq 6e^{-t}, \tag{36}$$

where

$$q(t) = 2\sqrt{\left(\frac{4\overline{\sigma}^2}{d_{\text{eff}}} + 16L\frac{\sqrt{2\overline{\sigma}^2}}{N}\right)t} + \frac{2L}{N}t. \tag{37}$$

Proof. Let us remark that

$$\left\langle \frac{1}{N}\sum_{k=1}^N X_k, \frac{1}{N}\sum_{k=1}^N Y_k \right\rangle = \frac{1}{2N^2}\left[\left\|\sum_{k=1}^N X_k + Y_k\right\|^2 - \left\|\sum_{k=1}^N X_k\right\|^2 - \left\|\sum_{k=1}^N Y_k\right\|^2\right].$$

So, by Corollary F.3, with probability greater than $1-6e^{-t}$, for $t \geq 1$, and using $(a-b)_+^2 \geq a^2 - 2ab$:

$$\begin{aligned}
2\left\langle \frac{1}{N}\sum_{k=1}^N X_k, \frac{1}{N}\sum_{k=1}^N Y_k \right\rangle &\leq \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N}} + q(t)\right)^2 - \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X}{N}} - 4q(t)\right)_+^2 \\
&\quad - \left(\sqrt{\frac{\operatorname{Tr}\Sigma_Y}{N}} - 4q(t)\right)_+^2 \\
&\leq q(t)(19\overline{\sigma} + q(t)) \leq 20q(t)\max(\overline{\sigma}, q(t)).
\end{aligned}$$

□

Proposition F.5. *Let $(X_k)_{1\leq i\leq N} \overset{i.i.d.}{\sim} X$ and $(Y_k)_{1\leq i\leq N} \overset{i.i.d.}{\sim} Y$ be independent families of random variables bounded by L in $\mathcal{H}$. Let μ_x, Σ_X and μ_Y, Σ_Y denote their respective means and covariance operators. Let U the statistic defined as*

$$U = \frac{1}{N(N-1)}\sum_{\substack{k,\ell=1\\k\neq\ell}}^N \langle X_k, X_\ell\rangle_{\mathcal{H}} - \frac{2}{N^2}\sum_{k,\ell=1}^N \langle X_k, Y_\ell\rangle_{\mathcal{H}} + \frac{1}{N(N-1)}\sum_{\substack{k,\ell=1\\k\neq\ell}}^N \langle Y_k, Y_\ell\rangle_{\mathcal{H}}. \tag{38}$$

Then for any $t \geq 1$, $N \geq 2$:

$$\mathbb{P}\Big[U \geq \|\mu_X - \mu_Y\|^2 + 2\|\mu_X - \mu_Y\| q(t) + 2\sqrt{2\overline{\sigma}^2} q(t) + 11 q^2(t)\Big] \leq 8e^{-t}, \tag{39}$$

and

$$\mathbb{P}\Big[U \leq \|\mu_X - \mu_Y\|^2 - 2\|\mu_X - \mu_Y\| q(t) - 8\sqrt{2\overline{\sigma}^2} q(t) - 32 q^2(t)\Big] \leq 8e^{-t}, \tag{40}$$

where $q(t)$ is given by (37).

Proof. Observe that

$$\begin{aligned} U &= \left\| \frac{1}{N}\sum_{k=1}^N X_k - \frac{1}{N}\sum_{k=1}^N Y_k \right\|^2 \\ &\quad + \frac{1}{N-1}\left(\left\| \frac{1}{N}\sum_{k=1}^N X_k \right\|^2 + \left\| \frac{1}{N}\sum_{k=1}^N Y_k \right\|^2 - \frac{1}{N}\sum_{k=1}^N \|X_k\|^2 - \frac{1}{N}\sum_{k=1}^N \|Y_k\|^2 \right) \\ &=: \left\| \frac{1}{N}\sum_{k=1}^N X_k - \frac{1}{N}\sum_{k=1}^N Y_k \right\|^2 + \frac{1}{N-1} H\,. \end{aligned}$$

Using now the upper bound of Bernstein's inequality, since $\mathbb{E}\big[\|X\|^2\big] = \|\mu_X\|^2 + \operatorname{Tr}\Sigma_X$, with probability greater than $1 - e^{-t}$ it holds:

$$\frac{1}{N}\sum_{k=1}^N \|X_k\|^2 \geq \operatorname{Tr}\Sigma_X + \|\mu_X\|^2 - \sqrt{2L^2\overline{\sigma}^2 t} - \frac{2L^2 t}{3N}\,.$$

So using (34) (twice), with probability greater than $1 - 6e^{-t}$:

$$\begin{aligned} H \leq& \|\mu_X\|^2 + 2\|\mu_X\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X}{N}} + q(t) \right)^2 + \|\mu_Y\|^2 + 2\|\mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_Y}{N}} + q(t) \right)^2 \\ &- \operatorname{Tr}\Sigma_X - \|\mu_X\|^2 + \sqrt{2L^2\overline{\sigma}^2 t} + \frac{2L^2 t}{3N} - \operatorname{Tr}\Sigma_Y - \|\mu_Y\|^2 + \sqrt{2L^2\overline{\sigma}^2 t} + \frac{2L^2 t}{3N} \\ \leq& -(N-1)/N\Big(\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y \Big) + 4Lq(t) + 4\sqrt{\overline{\sigma}^2} q(t) + 2q^2(t) + 2\sqrt{2L^2\overline{\sigma}^2 t} + \frac{4L^2 t}{3N} \\ \leq& -(N-1)/N\Big(\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y \Big) + (2+4N) q^2(t)\,. \end{aligned}$$

Using again (34), and $N \geq 2$, with probability greater than $1 - 8e^{-t}$:

$$\begin{aligned} U &\leq \|\mu_X - \mu_Y\|^2 + 2\|\mu_X - \mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N}} + q(t) \right)^2 - \frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N} + 10 q^2(t) \\ &\leq \|\mu_X - \mu_Y\|^2 + 2\|\mu_X - \mu_Y\| q(t) + 2\sqrt{2\overline{\sigma}^2} q(t) + 11 q^2(t)\,, \end{aligned}$$

which is (39).

We proceed similarly for lower deviations of U: using again Bernstein's inequality and (35), with probability greater than $1 - 6e^{-t}$, and using $(a-b)_+^2 \geq a^2 - 2ab$:

$$\begin{aligned} H \geq& \|\mu_X\|^2 - 2\|\mu_X\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X}{N}} - 4q(t) \right)_+^2 + \|\mu_Y\|^2 - 2\|\mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_Y}{N}} - 4q(t) \right)_+^2 \\ &- \operatorname{Tr}\Sigma_X - \|\mu_X\|^2 - \sqrt{2L^2\overline{\sigma}^2 t} - \frac{2L^2 t}{3N} - \operatorname{Tr}\Sigma_Y - \|\mu_Y\|^2 - \sqrt{2L^2\overline{\sigma}^2 t} - \frac{2L^2 t}{3N} \\ \geq& -(N-1)/N(\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y) - 16N q^2(t)\,, \end{aligned}$$

which implies, using again (35), and $N \geq 2$, that with probability greater than $1 - 8e^{-t}$ it holds:

$$U \geq \|\mu_X - \mu_Y\|^2 - 2\|\mu_X - \mu_Y\|q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N}} - 4q(t)\right)_+^2 - \frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N} - 16q^2(t)$$
$$\geq \|\mu_X - \mu_Y\|^2 - 2\|\mu_X - \mu_Y\|q(t) - 8\sqrt{2\overline{\sigma}^2}q(t) - 32q^2(t)\,,$$

which is (40).

□

G Details on the Tested Methods in the Numerical Experiments

In the following, the methods that are tested in the experiments are described in more detail. Recall, that $V_i := \{j : T_{ij} = 1, j \in [\![B]\!]\}$ and let T_{ij} be defined as in Eq (18), i.e. V_i holds the neighboring kernel means of bag i. All of the methods give KME estimations of the form

$$\tilde{\mu}_i := \sum_{j \in [\![B]\!]} \omega_{ij} \cdot \widehat{\mu}_j^{\mathrm{NE}},$$

where the definition of the weighting w_{ij} depends on the applied method.

1. `NE` considers each bag individually. Therefore, the weighting is simply

$$\omega_{ij} = \begin{cases} 1, & \text{for } i = j \\ 0, & \text{otherwise.} \end{cases}$$

2. `R-KMSE` was proposed by Muandet et al. (2016). It estimates each KME individually but shrinks it towards 0. The amount of shrinkage depends on the data and is defined as

$$\omega_{ij} = \begin{cases} 1 - \frac{\lambda}{1+\lambda}, & \text{for } i = j \\ 0, & \text{otherwise} \end{cases}$$

where

$$\lambda = \frac{\varrho - \rho}{(1/N_b - 1)\varrho + (N_b - 1)\rho}$$

with $\varrho = 1/N_i \sum_{k=1}^{N_i} k(Z_k^{(i)}, Z_k^{(i)})$ and $\rho = 1/N_i^2 \sum_{k,\ell=1}^{N_i} k(Z_k^{(i)}, Z_\ell^{(i)})$.

3. `STB-0` is described in Eq. (5) with γ set to 0, i.e.

$$\omega_{ij} = \begin{cases} \frac{1}{|V_i|}, & \text{for } j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

4. `STB theory` is defined by Eq. (5). It uses the optimal value for γ as described in Eq. (7) that was proven to be optimal. Here, τ is replaced by its empirical counterpart ζ and another multiplicative constant $c > 0$ was added to allow for more flexibility. Its specific value must be found using model optimization.

$$\omega_{ij} = \begin{cases} \gamma + \frac{1-\gamma_i}{|V_i|}, & \text{for } i = j \\ \frac{1-\gamma_i}{|V_i|}, & \text{for } i \neq j, j \in V_i \\ 0, & \text{otherwise} \end{cases}$$

with

$$\gamma_i = \frac{c \cdot \zeta \cdot (|V_i| - 1)}{(1 + c \cdot \zeta) \cdot (|V_i| - 1) + 1}.$$

5. `STB weight` is also described by Eq. (5) but the optimal value of γ is found by model optimization

$$\omega_{ij} = \begin{cases} \gamma + \frac{1-\gamma}{|V_i|}, & \text{for } i = j \\ \frac{1-\gamma}{|V_i|}, & \text{for } i \neq j, j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

6. `MTA const` is based on a multi-task averaging approach described in Feldman et al. (2014) which we translated to the KME framework as

$$\omega_{ij} = \left(\left(I + \frac{\gamma}{B} D \cdot L(A) \right)^{-1} \right)_{ij}. \tag{41}$$

   Here, $D = \operatorname{diag}\big((E_i)_{i \in [\![B]\!]}\big)$ as defined in Eq. (19) and $L(A)$ denotes the graph Laplacian of task-similarity matrix A. For `MTA const` the similarity is assumed to be constant, i.e. $A = a \cdot (\mathbf{1}\mathbf{1}^T)$ with $a = \frac{1}{B(B-1)} \sum_{i,j \in [\![B]\!]} \big\| \widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}} \big\|_{\mathcal{H}}^2$. Again, the optimal value for γ must be found using model optimization.

7. `MTA stb` is defined as in Eq. (41). In contrast to `MTA const`, the similarity matrix A is defined as

$$A_{ij} = \begin{cases} 1, & \text{for } j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

The methods `STB-0`, `STB weight`, `STB theory` and `MTA stb` use all the similarity test defined by T_{ij} which depends on ζ. Nevertheless, the optimal value for ζ is found by model optimization for each method individually.

H Numerical Results in the Gaussian Setting

In this section we report numerical comparisons of the proposed approaches in the idealized Gaussian setting (GI). In that setting, since the tests and proposed estimates only depend on the naive estimators, we can reduce each bag to its naive estimator, in other words we can assume $N = 1$ (only one observation per bag). We consider the following models for the means $(\mu_i)_{i \in [\![B]\!]}$ (in each case the number of bags is $B = 2000$):

- Model **UNIF**: ambient dimension $d = 1000$, the means $(\mu_i)_{i \in [\![B]\!]}$ are distributed uniformly over the lower-dimensional cube $[-20, 20]^{d'}$, $d' = 10$ (the remaining coordinates are set to 0).
- Model **CLUSTER**: ambient dimension $d = 1000$, the means are clustered in 20 clusters of centers $(m_i)_{i \in [\![10]\!]}$, drawn as $\mathcal{N}(0, I_d)$, in each cluster the means are drawn as Gaussians $\mathcal{N}(m_i, 0.1 * I_d)$,
- Model **SPHERE**: ambient dimension $d = 1000$, the 6 first coordinates of the means are distributed uniformly on the sphere of radius 50 in $\mathbb{R}^6$, the rest are set to 0.
- Model **SPARSE**: ambient dimension $d = 50$, the means are 2-sparse vectors with two random coordinates distributed as $\mathrm{Unif}[0, 20]$.

In each case, we first select the parameter for the tests (parameter ζ in (13)) from the oracle `STB-0` performance. This value is held fixed and the shrinkage parameter in methods `MTA stb`, `STB theory`, `STB weight` is again determined as its "oracle" value by minimization over the squared error, as done in the KME experiments.

For comparison, we also display the results of the classical positive-part James-Stein estimator (`PP James-Stein`, Baranchik, 1970), which is a shrinkage estimator applied separately on each bag. It has no tuning parameter.

Table 2: Decrease in averaged squared estimation error compared to `NE` in percent on the Gaussian data (higher is better). Averaged results over 20 trials. Standard error of one given trial is of order 5.10^{-3}.

	PP James-Stein	MTA const	MTA stb	STB-0	STB theory	STB weight
UNIF	0.439	0.427	0.653	0.796	0.813	0.813
CLUSTER	0.495	0.508	0.979	0.980	0.980	0.980
SPHERE	0.285	0.285	0.745	0.894	0.898	0.898
SPARSE	0.224	0.162	0.367	0.402	0.441	0.443

References

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 41(2):642–645.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 113–133. Inst. Math. Statist.

Blanchard, G., Carpentier, A., and Gutzeit, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electronic Journal of Statistics*, 12(2):3713–3735.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *Proc. of the 33rd International Conference on Machine Learning (ICML 2016)*, volume 48, pages 2606–2615.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637.

Fathi, M., Goldstein, L., Reinert, G., and Saumard, A. (2020). Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. arXiv preprint 2004.01378.

Feldman, S., Gupta, M. R., and Frigyik, B. A. (2014). Revisiting Stein's paradox: multi-task averaging. *Journal of Machine Learning Research*, 15(106):3621–3662.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379. University of California Press.

Jegelka, S., Gretton, A., Schölkopf, B., Sriperumbudur, B. K., and Von Luxburg, U. (2009). Generalized clustering via kernel embeddings. In *Annual Conference on Artificial Intelligence (KI 2009)*, pages 144–152. Springer.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer.

Martínez-Rego, D. and Pontil, M. (2013). Multi-task averaging via task clustering. In *Proc. Similarity-Based Pattern Recognition - Second International Workshop, SIMBAD 2013*, pages 148–159.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture notes in mathematics*. Springer.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2012). Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pages 10–18.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. (2016). Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17(48):1–41.

Pinelis, I. and Sakhanenko, A. I. (1986). Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148.

Rio, E. (2002). Une inégalité de Bennett pour les maxima de processus empiriques. *Annales de l'IHP Probabilités et statistiques*, 38(6):1053–1057.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Proc. International Conference on Algorithmic Learning Theory (ALT 2007)*, pages 13–31.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, volume 1, pages 197–206. University of California Press.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126:505–563.

Wang, Z., Lan, L., and Vucetic, S. (2011). Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2226–2237.